American Sign Language Recognition Using Convolutional Neural Networks

Fatima-Zahrae El-Qoraychy
UTBM, CIAD UMR 7533,
F-90010 Belfort cedex, France
email:fatima.el-qoraychy@utbm.fr

Yazan Mualla
UTBM, CIAD UMR 7533,
F-90010 Belfort cedex, France
email:yazan.mualla@utbm.fr

***Abstract*—Sign Language Recognition (SLR) poses a challenge due to the rapid and intricately coordinated motions inherent in gestures. This research endeavors to address this complexity by leveraging Convolutional Neural Networks (CNNs). It presents a comprehensive exploration of diverse studies, methodologies, and inherent challenges in SLR, with a specific focus on harnessing CNN-based approaches for enhanced comprehension. At the core of this study lies a project aimed at the classification of American Sign Language gestures using CNN models rooted in the Visual Geometry Group 19 architecture. This initiative seeks to enrich the understanding and interpretation of manual gestures, fundamental to effective communication. Within this context, the article delves into pivotal aspects encompassing data diversification, model performance, and prospective limitations. Practical remedies are proposed, including data set augmentation and the incorporation of image masks, with the explicit objective of fortifying the precision and robustness of gesture recognition. For the validation and elucidation of classification outcomes, this study integrates the Gradient-weighted Class Activation Mapping (Grad-CAM) explanation model. This model uncovers salient regions within images, shedding light on the decision-making mechanisms of the CNN model, thereby enhancing transparency and comprehension.**

Keywords-Human-Computer Interaction; Convolutional Neural Networks; Sign Language Recognition

I. INTRODUCTION

In light of accelerated progress in Artificial Intelligence (AI) and its integration across multifaceted aspects of human existence, the interaction between humans and technological systems has assumed salience. The field of Human-Computer Interaction (HCI) focuses on the exchange of information and commands between human users and technological systems or computer devices. Many terms are used to represent the technology that the human interacts with, including computer, machine, AI, agent, robot. In the same vein, many relations could take place including interaction, cooperation, collaboration, team, symbiosis, integration [1]. This interaction takes diverse forms, such as text input, voice commands, gestures, eye movements, etc., and is ubiquitous in our daily lives, from smartphones and computers to cars and robots. Enhancing these interactions is crucial to making technological systems more user-friendly, efficient, and tailored to users' needs.

In recent years, notably within the past decade, the field of Sign Language Recognition (SLR) has witnessed significant advancements, thanks to the application of AI and Computer Vision techniques. SLR plays a pivotal role in facilitating communication between the Deaf or Hard-of-Hearing community and the hearing population. It accomplishes this by interpreting sign language gestures, converting them into text or speech, and effectively bridging the communication gap to enable seamless interaction between individuals with different language modalities.

Recent years have witnessed the emergence of CNNs as potent tools for image and video-based recognition tasks, notably within the realm of SLR. These deep learning models have demonstrated remarkable performance in recognizing both static and dynamic sign gestures from video sequences or individual frames. However, despite the success of CNNs in SLR, several challenges need to be addressed to enhance the overall effectiveness of SLR systems.

The primary focus of this scientific article is adopting and adapting the existing model from the American Sign Language (ASL) project. The current SLR system relies on a CNN-based model, trained on a substantial dataset containing various sign gestures captured as images. While the model exhibits satisfactory accuracy on the training set, it still faces difficulties in recognizing complex gestures and identifying specific gestures during testing. Additionally, it lacks transparency in decision-making and presents limitations in adapting to regional and individual sign variations.

To overcome these limitations and enhance the performance of the SLR system, we are exploring the integration of Explainable Artificial Intelligence (XAI) techniques to validate the obtained results and improve the module's performance. XAI has emerged to enhance the interaction between humans and computers. XAI refers to the ability of an AI system to provide clear and understandable explanations for its decisions and actions. This functionality is crucial to understanding how and why an AI makes a specific decision, enabling the evaluation of its reliability, identification of potential biases, resolution of trust issues, and ensuring the ethical use of AI systems. Acknowledging the significance of XAI, the Defense Advanced Research Projects Agency (DARPA) initiated the "XAI Program" in 2017 [2], which propelled research into enhancing AI explainability. This momentum has yielded noteworthy contributions: The HAExA architecture [3] furnishes lucid agent decision explanations. "DExAI: Driving-X" [4] offers neural network action insights in autonomous vehicles. "DExAI: Saliency Driven Retrieval" [5] improves

ISBN: 978-1-68558-100-8

image search via saliency maps. RISE [6] generates neural network importance maps. These projects aim to showcase the use of XAI to explain and interpret artificial intelligence. XAI is designed to enhance understanding for both users and the machine. Additionally, its role is to elucidate the interaction between humans and machines. In this perspective, our project focuses on this domain.

This paper aims to contribute to the field of HCI by enhancing an existing ASL project. The article's structure is organized as follows: In Section II, we provide an overview of the fundamentals of SLR. Section III presents the existing ASL project and its current limitations. Section IV outlines our proposed approach to enhance the SLR system, emphasizing the incorporation of XAI techniques and data enrichment strategies, and covers the evaluation metrics used to assess the improved SLR model's performance. Finally, Section V concludes the article with a summary of our contributions and potential future research directions.

II. Related Work

The domain of SLR has witnessed remarkable progress, driven by the growing need for inclusive communication within the deaf and hard-of-hearing communities. Recent years have seen significant advancements in Deep Learning (DL) methodologies, synergistically contributing to enhanced accuracy in SLR systems. This section presents an in-depth review of relevant scholarly literature, focusing specifically on CNN-based paradigms.

An exemplary contribution in the realm of SLR is exemplified by Kumar et al. [7]. The mentioned study introduces a communication system designed to assist individuals with vocal and hearing impairments. The system employs skin color segmentation to extract sign language gestures from videos, utilizing a CNN to learn and classify visual features associated with these gestures. Additionally, the system utilizes the Sphinx module to recognize spoken language and convert it into corresponding sign language gestures.

The endeavor by Devineau et al. [8] is equally noteworthy. The study presents a novel approach to hand gesture recognition using deep learning and skeletal data. The authors use a CNN to learn features from the 3D coordinates of the hand joints captured by a depth sensor. The CNN is trained on a large dataset of 14 hand gestures performed by 28 subjects. The experimental results show that the proposed method achieves high accuracy and robustness in recognizing hand gestures, outperforming existing RGB or depth images. The article demonstrates the potential of using skeletal data as a low-dimensional and noise-resistant representation for hand gesture recognition.

The landscape of CNN-based SLR is further illuminated by the work of DeVries et al. [9]. This scholarly exposition introduces a tailored CNN-driven framework designed for SLR. Notably, this framework navigates the multifaceted challenges stemming from the intrinsic variability of hand gestures. Within the mentioned work, innovative solutions are proposed, with the overarching goal of enhancing model efficacy and performance. Furthermore, the pursuit of real-time applications within SLR is exemplified by Garcia et al. [10]. The authors design a custom CNN model that can process video frames of hand gestures and output the corresponding ASL letters. The model is trained and tested on a large dataset of 24 ASL letters performed by 10 subjects. The experimental results show that the proposed architecture achieves high accuracy and speed in recognizing ASL letters, outperforming existing methods that use hand-crafted features or other deep learning models. The article demonstrates the feasibility and effectiveness of using CNNs for real-time ASL recognition.

In summary, this comprehensive collection of scholarly endeavors underscores the evolutionary trajectory of SLR through the lens of CNN-based approaches.

III. The existing ASL Project

In the scope of this work, the central focus is on the classification of ASL gestures, a fundamental step for the understanding and interpretation of sign language. Recognizing the inherent complexity of such systems, a current model of ASL classification presents intrinsic limitations. Consequently, this research strives to expand the current boundaries by meticulously identifying and addressing these constraints through targeted methodologies. The primary objective is to refine the model's understanding and enhance its overall performance. To materialize this ambition, a range of meticulously designed solutions is proposed to alleviate the identified limitations.

A. Overview

To interpret and classify ASL gestures, we enhanced a project initiated by Damion Joyner. [11] that aims to classify a set of RGB and depth images of ASL using a CNN model based on the Visual Geometry Group 19 (VGG19) architecture. The model is trained using the ASL alphabet dataset [12]. This dataset comprises over 100,000 images of English alphabet letters in sign language from 5 different individuals. Given that there are 24 letters in the English alphabet (excluding the letters 'g' and 'z' as they require hand movement) and the images are provided by 5 pairs of hands, the model must be capable of classifying images based on the different letters. To comprehend the functionalities of this model, along with the achieved results and potential enhancements, and since the classification model based on VGG19 is necessary to understand the VGG architecture, starting with VGG19. This architecture plays a pivotal role in constructing and training the ASL hand gesture classification model.

B. Classification Model Structure

1) Visual Geometry Group: Visual Geometry Group (VGG) is a standard CNN architecture known for its depth, signifying the high number of convolutional layers it comprises. VGG has been instrumental in pioneering object recognition models, surpassing benchmarks in numerous tasks and datasets. Even today, VGG remains one of the most popular image recognition architectures [13]. VGG19, proposed by Simonyan and Zisserman [14], is an enhanced version of the VGG architecture with 19 convolutional layers. It consists of several convolutional blocks, each comprising multiple convolutional layers followed by pooling layers. The model utilizes small-sized filters (3×3) with a pattern of stride of 1 and padding of 1 to preserve extracted feature sizes. Using this pattern for the convolutional layers means that the convolutional filters move one pixel at a time across the input data, and one layer of zero pixels is added around the input to maintain its size during the convolution process. After the convolutional blocks, the network connects to fully connected layers for classification. The classification model presented by Damion Joyner [11] is a combination of the pre-trained VGG19 model and additional layers added for the specific task of image classification. Here is an overview of the breakdown between the VGG19 layers and the added layers:

a) VGG19 Model Layers:: The VGG19 layers follow the standard architecture of VGG19, including blocks of convolutional layers followed by pooling layers. These layers progressively capture features at different scales and complexities.

b) Additional Layers:: Several layers are added after the VGG19 layers to adapt the model for the image classification task. These additional layers include:

- A flattened layer to transform the outputs into a one-dimensional vector.
- Dense (fully connected) layers for final classification. These layers include dropout layers for regularization.
- Batch normalization layers for normalizing activations and stabilizing learning.
- The final dense layer, with neurons corresponding to the number of classes (letters) in the classification problem.

The classification model is used for both RGB and depth images. Initially, the model was applied individually to each type of data, resulting in separate classification models. Subsequently, the model was trained on the combined dataset of RGB and depth images to explore the potential benefits of multi-modal learning.

2) Model Performance and Limitations: The classification model has exhibited remarkable performance, with accuracy exceeding 95% on the test dataset, effectively showcasing its ability to forecast the English letters corresponding to the gestures precisely. However, it is worth noting that in the author's project test [11], the model faced difficulties in correctly predicting the class for each hand gesture, indicating a limitation that persists in the model. Despite its promising performance, the classification model does exhibit certain limitations:

- Lack of Diversity: The ASL alphabet dataset primarily consists of images from 5 individuals, potentially limiting the model's ability to generalize to a broader population.
- Overfitting: The model might suffer from overfitting, especially considering the dataset's limited size and potential data imbalances.
- Multimodal Integration: While the model was trained on combined RGB and depth data, there is potential for further exploring how to effectively integrate information from different modalities.

3) Proposed Techniques: Addressing the limitations is crucial for enhancing the classification model's performance and robustness. In the following sections, we discuss our potential solutions and strategies to mitigate these limitations:

- **Data Augmentation and Diversification**

To address the challenge of limited diversity in the dataset, we suggest the implementation of data augmentation techniques. By applying transformations, such as rotations, flips, and adjustments to brightness, the augmentation process can be further enhanced by collecting additional images from a variety of hand sources. This approach aims to enrich the dataset, exposing the model to a wider range of hand shapes and features. Consequently, the model's capacity to generalize and recognize signs performed by different individuals can be significantly improved. The collected dataset consists of cropped RGB images depicting ASL hand shapes corresponding to the 26 letters of the English alphabet. Instead of utilizing 100,000 images, we employ 436,433 images to enhance the dataset's richness and diversity.

The image data utilized in our work has been sourced from various origins, including:

- Kaggle - ASL Alphabet [15]
- Kaggle - ASL RGB Depth Finger spelling [12]
- Kaggle - ASL American Sign Language Alphabet Dataset [16]
- Kaggle - ASL Alphabet Test [17]
- Kaggle - Synthetic ASL Alphabet [18]

These diverse data sources contribute a wide array of images, representing distinct letters of the ASL alphabet.

- **Mask Image Approach Instead of Depth Images**

This approach proposes substituting depth images with image masks to create a more effective representation of ASL gestures. Rather than relying on raw depth data, the concept involves using masks to accentuate the critical areas of gestures, specifically, the regions where hand movements occur. By leveraging masks, we can accentuate the essential intricacies of the gestures while excluding background elements. This strategy has the potential to minimize data noise and concentrate on the distinct characteristics of ASL gestures, thereby enhancing the model's capacity to generalize and discriminate between various letters. For this purpose, the acquisition of an image segmentation dataset is necessary. The dataset we have come across is HGR1 [19], containing 899 images. Initially tailored for recognizing diverse signs in both Polish and ASL, this dataset can be conveniently adapted for alternative applications. Comprising images of hands from various individuals, the dataset encompasses a range of backgrounds, varying lighting conditions, and diverse capture angles. It also provides hand segmentation masks. The images in this dataset showcase different proportions, sizes, and resolutions, as they were captured using an assortment of cameras.

IV. Our proposed Solution

The realization of this work unfolds in three essential steps, each contributing to the achievement of our ultimate goal. The first step involves image mask extraction, where we apply image processing techniques to isolate hand regions in the captured images. This step serves to reduce noise and focus on the relevant parts for gesture classification. The second step is the training of the letter prediction model. We utilized the model presented in the existing project, to train our model on tailored datasets. This step is crucial to harness the visual features of hand gestures and enable accurate classification of ASL letters. The third and final step of our work involves the use of an explainable AI model to validate our prediction model. Explainability is a crucial feature to ensure users' confidence and acceptance of AI systems. Figure 1 presents the architecture of our solution.

A. Segmentation Model

A pivotal step for accurate gesture recognition is hand segmentation. Hand segmentation is a highly active research domain [20]. The primary goal of hand segmentation is to identify the pixels composing the hands in an image and represent them as a mask. Once the mask is obtained, various analyses can be performed, such as separating the hands from the background or further analysis [21]. Numerous methods exist for performing hand segmentation, including skin color analysis, machine learning-based modeling, and more [22]. In this work, we use U-NET, a deep learning-based method widely acclaimed for image segmentation, particularly in medical imagery [23]. The name "U-NE" is inspired by its architectural shape, resembling the letter "U". This unique design involves connecting the outputs of corresponding layers both above and below the U shape. Essentially, these outputs directly link to other filters in the convolutional layers, forming a U-shaped structure.

To maintain consistent training image dimensions, we standardize the image resolution across all training data. Following this, we split the dataset into training and testing segments. Moving forward, we employ data augmentation using rotated images, ensuring caution in transformations, particularly concerning skin color. We refrain from altering the color and avoid excessive deformation, recognizing the distinctive shape of hands that our model must recognize. Applying the U-NET method to our dataset, we extract masks from the images to train the gesture recognition model.

B. Model for Gesture Recognition

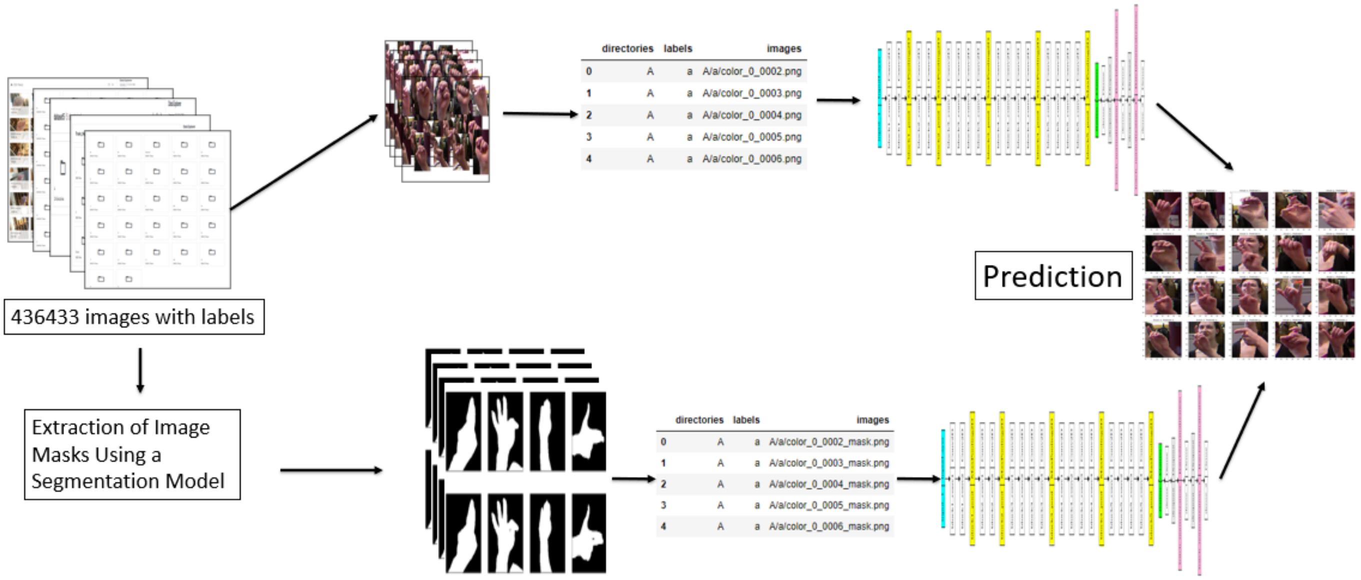
After extracting the image masks using the U-Net model, we will now train two classification models using the existing project's classification model presented in Section III-B. As shown in Figure 1, each model will be trained on a different dataset. The first model will be trained on the image masks, while the second will be trained on RGB images.

The results obtained after training the two models have demonstrated exceptional performance in accurately categorizing a wide variety of hand gesture images. The model evaluation revealed high and consistent precision, recall, and F1 scores for multiple classes. Specifically, precision scores ranged from 0.81 to 0.99 for each class, reflecting the model's ability to make highly accurate predictions. Similarly, recalls ranged from 0.86 to 0.99 for each class, highlighting the models' ability to identify instances of different classes.

The idea of using two classification models—one on the image masks and the other on real images—and then combining their outputs to obtain the exact classification has proven successful. After training both models, the results are very satisfactory for both versions. However, there are some differences between the two models. The model trained on RGB images demonstrates adeptness in accurately detecting all test images with high precision, even in scenarios where hand gestures are similar, effectively identifying the corresponding letter. On the other hand, the mask-based model occasionally makes errors, particularly when distinguishing between similar gestures such as the letters ”A” and ”E.” Figures [2, 3] illustrate the results of precision, F1-score, and recall for both models. The diagram for the mask model displays lower results than the RGB model, primarily attributed to challenges in detecting similar gestures. To gain insight into how the model makes decisions and to make a comparison between the two models, an explanatory model becomes essential for visualizing image components that influence predictions. Therefore, the incorporation of an explanatory model is imperative for a comprehensive understanding.

C. Explanation Model

The explanation model utilized is the Grad-CAM model, which was integrated into the classification model to validate the classification outcomes. Grad-CAM aids in comprehending the image regions that significantly influenced the classification decision made by the model. The generation of an activation map highlights the portions of the image that played a positive or negative role in shaping the model's prediction. Employing the Grad-CAM model in conjunction with the classification model allows us to visually interpret the regions of interest leveraged by the model in reaching its classification verdict. This insight permits verification of whether the model is focusing on the hand and provides insight into the logic underpinning specific decisions. The Grad-CAM algorithm yields a heatmap as its output, accentuating the image regions that contributed most to the classification prediction of the target class by the model. The heatmap assigns weights to different regions of the image, thereby indicating their relative importance. The coloring scheme employed in the heatmap varies based on the chosen color map. The "JET" color map is utilized in this work, commonly employed for heatmap visualization. In this color map, warmer regions are represented by vivid colors like red, orange, and yellow, while cooler regions are depicted by shades of blue and violet. Consequently, within the heatmap, regions tinted in red, orange, and yellow signify


436433 images with labels
Extraction of Image Masks Using a Segmentation Model
Prediction

	directories	labels	images
0	A	a	A/a/color_0_0002.png
1	A	a	A/a/color_0_0003.png
2	A	a	A/a/color_0_0004.png
3	A	a	A/a/color_0_0005.png
4	A	a	A/a/color_0_0006.png

	directories	labels	images
0	A	a	A/a/color_0_0002_mask.png
1	A	a	A/a/color_0_0003_mask.png
2	A	a	A/a/color_0_0004_mask.png
3	A	a	A/a/color_0_0005_mask.png
4	A	a	A/a/color_0_0006_mask.png

Figure 1. The adapted solution

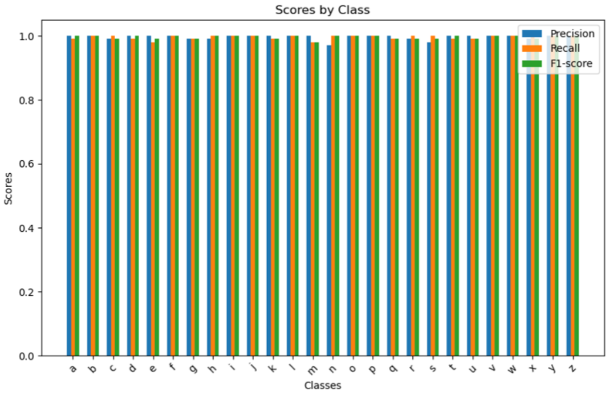


Figure 2. Result RGB Classification Model

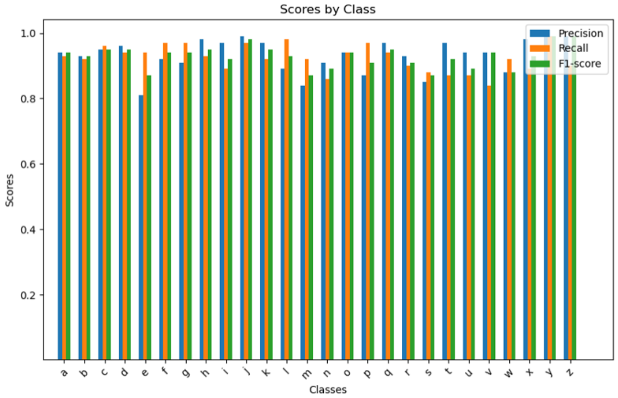


Figure 3. Result Mask Classification Model

the most pivotal areas governing predictions for the target class. Conversely, regions shaded in blue and violet denote areas of lesser significance. We will apply Grad-CAM to both classification models to understand the regions on which the model relies to make its decision. To do this, we will choose a test image. The selected image contains the letter “A” Figure 4. The models successfully detected the image’s class. Now, we will determine which region of the image enabled this decision. Let’s start with the model trained on RGB images. As illustrated in Figure 5, the Grad-CAM model can identify the hand throughout the entire image. This indicates that the predictions of the classification model, trained on RGB images, rely on information from the entire image. When applying Grad-CAM to the model trained on image masks (see Figure 6), the results depicted in Figure 7 reveal that the region primarily influencing the decision is the hand. Consequently, the model primarily focuses on the hand to make predictions, which is logical given that the image only contains the hand mask. Therefore, we can conclude that the model trained with mask images is more effective than the RGB model because the predictions are based on the hand, which is the most important feature.

V. Conclusion

This article aims to explore the potential of enhancing SLR through the application of advanced AI techniques. Focusing on hand image segmentation and gesture classification, we enhance existing projects that employ approaches, such as the VGG19 model, and we add the CNN U-NET method to achieve promising results.

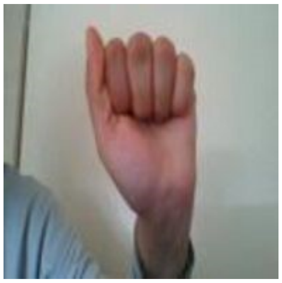


Figure 4. Test image, RGB Classification Model

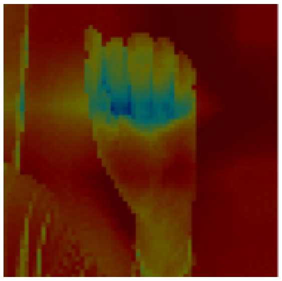


Figure 5. Grad-CAM Visualization, RGB Model

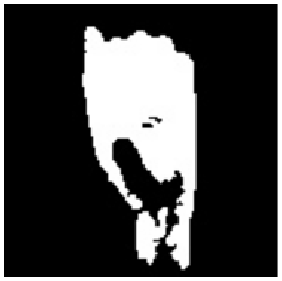


Figure 6. Mask image, mask Classification Model

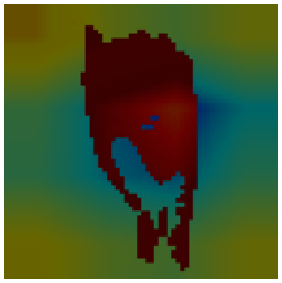


Figure 7. Grad-CAM Visualization, mask Model

The examination of previous work in the field of gesture recognition has underscored the importance of robust and explainable models for effective and socially relevant applications. By integrating explainability methods, such as the Grad-CAM model, we were able to not only achieve accurate classifications but also comprehend the areas of interest guiding these classifications. Despite the successes encountered, it is essential to acknowledge the limitations of our approach, particularly in terms of data diversity and the risks of overfitting. These challenges pave the way for future research aimed at improving performance, expanding the scope of the approach to more diverse populations, and exploring other data modalities. Additionally, the use of other explainable models to enhance the explanation and interpretation of the project is recommended. This work highlights the potential of AI to enhance communication and accessibility for individuals using ASL. We hope that our findings will encourage other researchers to continue in this direction, developing more sophisticated approaches, exploring new data modalities, and contributing to broader inclusion and a better understanding of gestures in society.

Ultimately, this research demonstrates the positive impact that emerging technologies, combined with a deep understanding of the field, can have on individuals' daily lives and interactions. By combining the power of AI with the intricacies of gesture recognition, we aspire to have laid the foundation for improved communication and increased inclusion for individuals using sign language.

REFERENCES

[1] A. Picard, Y. Mualla, F. Gechter, and S. Galland, "Human-computer interaction and explainability: Intersection and terminology," In The World Conference on eXplainable Artificial Intelligence, July 2023;pp. 214–236.

[2] D. Gunning, "Explainable artificial intelligence (xai)," Defense Advanced Research Projects Agency (DARPA), 2017;pp. 1.

[3] Y. Mualla, I. Tchappi, T. Kampik, A. Najjar, D. Calvaresi, A. Abbas-Turki, S. Galland, and C. Nicolle, "The quest of parsimonious XAI: A human-agent architecture for explanation formulation," Artif. Intell. Vol. 302, 2022; pp. 103–573.

[4] University of California, Berkeley (UCB), "Deeply explainable artificial intelligence (dexai): Driving-x," https://www.darpa.mil/attachments/XAIProgramPortfolio.pdf, 2019.

[5] D. Darrell, T. Collins, and R. Roddy, "Deeply explainable artificial intelligence (dexai): Saliency driven retrieval," https://www.darpa.mil/attachments/XAIProgramPortfolio.pdf, 2019.

[6] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized Input Sampling for Explanation of Black-box Models," In Proceedings of the British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, 3–6 September 2018; BMVA Press: Durham, UK, 2018;p. 151.

[7] A. Kumar, K. Thankachan, and M. Dominic, "Sign language recognition," International Conference on Recent Advances in Information Technology (RAIT), 2016;pp. 422–428.

[8] G. Devineau, F. Moutarde, W. Xi, and J. Yang, "Deep learning for hand gesture recognition on skeletal data," International Conference on Automatic Face Gesture Recognition, 2018; pp. 106–113.

[9] L. Pigou, S. Dieleman, P. Kindermans, and B. Schrauwen, "Sign language recognition using convolutional neural networks," In Computer Vision ECCV, 2014;pp. 572–578.

[10] B. Garcia and S.A. Viesca, "Real-time american sign language recognition with convolutional neural networks,",2016;pp. 225-232.

[11] D. Joyner, "Sign-language-classification-cnn-vgg19," https://www.kaggle.com/code/damionjoyner/sign-language-classification-cnn-vgg19. [retrieved: November, 2023].

[12] N. Pugeault and R. Bowden, "Spelling it out: Real-time ASL fingerspelling recognition," IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain,2011; pp. 1114-1119.

[13] G. Boesch, "Vgg very deep convolutional networks(vggnet)," https://viso.ai/deep-learning/vgg-very-deep-convolutional-networks/. [retrieved: November, 2023].

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," CoRR, 2015, arXiv 1409.1556.

[15] Akash, "Asl alphabet,"https://www.kaggle.com/datasets/grassknoted/asl-alphabet. [retrieved: November, 2023].

[16] D. Sau, "Asl american sign language alphabet dataset," https://www.kaggle.com/datasets/debashishsau/aslamerican-sign-language-aplhabet-dataset. [retrieved: November, 2023].

[17] D. Rasband, "Asl alphabet test," https://www.kaggle.com/datasets/danrasband/asl-alphabet-test. [retrieved: November, 2023].

[18] Lexset, "Synthetic asl alphabet," https://www.kaggle.com/datasets/lexset/synthetic-asl-alphabet. [retrieved: November, 2023].

[19] M. Kawulok, "Database for hand gesture recognition," https://sun.aei.polsl.pl//m̃kawulok/gestures/.[retrieved: November,2023].

[20] Karen Mosoyan. "Hand segmentation with python and tensor-flow," https://medium.com/@karen.mossoyan/hand-segmentation-with-python-and-tensorflow-70c38db855b5. [retrieved: November, 2023].

[21] Z. Chen, J.-T. Kim, J. Liang, J. Zhang, and Y.-B. Yuan, "Real-time hand gesture recognition using finger segmentation," The Scientific World Journal, 2014.

[22] M. Ben Abdallah, A. Sessi, M. Kallel, and M. Bouhlel, "Different techniques of hand segmentation in the real time," Ijcait, 2013; pp.45–49.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation,". In : Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015;pp. 234-241.